Virology 450-451 (2014) 165–173

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Conserved features of an RNA promoter for RNA polymerase II determined from sequence heterogeneity of a hepatitis delta virus population

Yasnee Beeharry, Lynda Rocheleau, Martin Pelchat*

Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, 451 Smyth Road, Room 4111 A, Ottawa, Ontario, Canada, K1H 8M5

ARTICLE INFO

Available online 1 January 2014

Keywords:
RNA covariation
High-throughput sequencing
Secondary structure alignment
Viral quasi-species
Hepatitis *delta* virus
Sequence analysis

ABSTRACT

The right terminal domain of genomic hepatitis *delta* virus (HDV) RNA is involved in viral replication by recruiting host RNA polymerase II. To identify conserved features of this region, we performed high-throughput 454 sequencing of an HDV population actively replicating in cells. We generated 473,139 sequences representing 2351 new HDV variants of this region and investigated nucleotide conservation and positions of covariation in the population. We found that the sequence is heterogeneous and the rod-like conformation is conserved for both polarities of the HDV RNA genome at this location. Additionally, we identified conserved nucleotides near the previously reported initiation site of transcription, and corroborated our finding with sequences from HDV variants isolated in various hosts. Our analysis highlights the importance of both a conserved sequence at the tip of the rod-like structure and the RNA secondary structure upstream of the tip, which are likely important for HDV replication.

Introduction

A remarkable feature of most RNA viruses is that following replication they form a heterogeneous population of sequences (Borderia et al., 2011; Domingo et al., 2012). These differences are thought to result from infidelity of the replication machinery, which generates a heterogeneous population of RNA species within and among hosts, some of which are functional or even ameliorated, and can accumulate to become the most abundant species (Borderia et al., 2011; Domingo et al., 2012). Analogous to other RNA viruses, sequencing of a few variants of the hepatitis *delta* virus (HDV) indicated that this RNA virus accumulates mutations during its replication (Wang et al., 1986; Chao et al., 1990, 1991), and folding of its RNA genome into precise structures is considered to be required for its replication (Beard et al., 1996; Wu et al., 1997; Gudima et al., 1999; Filipovska and Konarska, 2000; Greco-Stewart et al., 2007; Abrahem and Pelchat, 2008).

HDV is one of the smallest animal virus known, and requires the hepatitis B virus (HBV) envelope proteins for its propagation (reviewed by (Taylor, 2009)). Its genome is composed of a single-stranded, circular RNA molecule of approximately 1700 nucleotides (nt), and folds into an unbranched, rod-shaped structure due to ~70% self-complementarity (Fig. 1). HDV contains one open reading frame (ORF) allowing for the synthesis of two viral proteins due to editing of antigenomic HDV RNA at the location of the termination codon of the small *delta* antigen (HDAg-S) ORF (Casey, 2012). HDAg-S (195 amino acids) is essential for HDV accumulation (Kuo et al., 1989; Yamaguchi et al., 2001). The large HDAg (HDAg-L; 214 amino acids) contains 19 additional amino acids at its C-terminus and is required for virion assembly (Chang et al., 1991; Ryu et al., 1992; Sureau et al., 1992).

HDV replicates through a symmetrical rolling circle mechanism that involves only RNA intermediates (reviewed by (Taylor, 2009)). Circular genomic strands of HDV RNA are used as templates for the synthesis of both the HDV mRNA and the linear multimeric anti-genomic strands. The latter are cleaved in monomers by self-cleaving motifs encoded on the left terminal region of both genomic and antigenomic HDV RNA polarities (Fig. 1; Rz motifs). The mechanism leading to the ligation and circularization of the genome is still unknown. Using the same steps, antigenomic strands are used as templates to generate the genomic RNA, which is found at a greater intracellular abundance than the antigenomic species.

HDV does not encode its own replicase and can replicate independently from HBV. The HDV RNA genome is capable of redirecting host DNA-dependent RNA polymerase II (RNAP II) for its own replication and transcription, although the manner by which this template switching from DNA to RNA occurs remains largely unknown. HDV RNA accumulation is sensitive to low levels of α-amanitin, a mycotoxin that inhibits DNA-dependent RNA synthesis by RNAP II (MacNaughton et al., 1991; Fu and Taylor, 1993; Filipovska and Konarska, 2000; Moraleda and Taylor, 2001; Chang et al., 2008).

* Corresponding author. Fax: +613 562 5452.
E-mail address: mpelchat@uottawa.ca (M. Pelchat).

0042-6822/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.virol.2013.12.017

Fig. 1. Schematic representation of the HDV RNA genome. The circular rod-like structure of the HDV genome is illustrated. The white arrow indicates the location of the HDAg ORF (i.e. initiating at nucleotide 1598). The Rz-G and Rz-AG boxes indicate the locations of both genomic and antigenomic polarities of the HDV ribozymes, respectively. A black arrow indicates the reported site of initiation of transcription from genomic HDV RNA (Beard et al., 1996; Gudima et al., 2000; Abrahem and Pelchat, 2008). The right terminal domain of genomic HDV RNA involved in viral replication is indicated by the gray rectangle (i.e. nucleotides 1541 to 60 of the genomic polarity). The sequences analyzed in this study correspond to nucleotides 1566 to 18 of the genomic polarity (i.e. between the locations indicated by the inverted fonts). The numbering is in accordance with (Kuo et al., 1988).

Several studies indicated a role for the right terminal domain of genomic HDV RNA in viral replication (Fig. 1; gray rectangle). In infected cells, the 5′ end of HDAg mRNA localizes in this region (i.e. position 1630; arrow on Fig. 1; (Gudima et al., 2000)). HDAg mRNA is post-transcriptionally processed with a 5′-cap and a 3′-poly (A) tail (Gudima et al., 1999, 2000), which suggests RNAP II involvement in the production of this mRNA. In vitro, this region acts as template to initiate antigenomic RNA synthesis, and the transcription reaction is inhibited by an antibody raised against the C-terminal domain of RNAP II (Abrahem and Pelchat, 2008). RNAP II forms an active pre-initiation complex on the right terminal domain of genomic HDV RNA, and this complex contains the same general transcription factors as those found on a typical DNA promoter (Greco-Stewart et al., 2007; Abrahem and Pelchat, 2008). The TATA-binding protein, alone or within the TFIID complex, directly binds the RNA promoter, and was proposed to be required to nucleate the RNAP II pre-initiation complex (Abrahem and Pelchat, 2008). Notably, mutations affecting the secondary structure of this region were reported to decrease HDV RNA accumulation in cell culture (Beard et al., 1996; Wu et al., 1997; Gudima et al., 1999), lower RNAP II affinity (Greco-Stewart et al., 2007; Abrahem and Pelchat, 2008), and affect RNAP II transcription initiation efficiency in vitro (Beard et al., 1996; Abrahem and Pelchat, 2008).

Although many studies have investigated DNA promoter recognition by RNAP II, little is known regarding how this enzyme recognizes an RNA template. Analogous to what is observed on DNA promoters (reviewed by (Baumann et al., 2010)), essential HDV RNA features (i.e. sequence and secondary structure) should be conserved and selected for during viral replication. By taking advantage of next-generation sequencing technology, the goal of this study was to investigate the positions of covariation and nucleotide conservation in a large population of heterogeneous HDV RNA sequences to define the selected features on the right terminal domain of genomic HDV RNA. We generated 2351 new HDV variants of this region derived from 473,139 sequences obtained by high-throughput 454 sequencing and originating from an HDV population replicating in a cellular system. We developed a pipeline to filter, align and analyze sequence conservation and covariation of this region from the population of sequences. Our results indicated the polymorphic nature of this segment of HDV, by showing that it accumulates as a population of different sequences. Despite sequence heterogeneity, our analyses revealed the conservation of the rod-like conformation of this region and identified conserved nucleotides at the tip of the rod-like structure, near the proposed transcription initiation site. These conserved features, which are also found on sequences from HDV variants isolated from various hosts, are likely important for HDV replication by RNAP II, and will be useful at identifying other RNA promoters for RNAP II.

Results

High-throughput sequencing of the right terminal region of genomic HDV RNA from a cellular system

To investigate the features of the right terminal domain of genomic HDV RNA involved in replication (Fig. 1; gray rectangle), we needed a system where the selective pressure was mainly on viral replication. We used the HDV replication system previously developed by Chang et al. (2005). In this system, 293 cells contain a replicating HDV RNA genome with a frame-shift deletion in the HDAg ORF, and allow the synthesis of HDAg-S under the control of a promoter inducible by tetracycline (Chang et al., 2005). Because a low level of HDAg-S is produced in the cells without induction, basal HDV replication is possible for several months, and HDV RNA genomes capable of replication are amplified upon tetracycline induction (Chang et al., 2005).

The 293-HDV cells were grown for more than a year without induction of HDAg-S expression to allow the accumulation of mutations on the HDV RNA genome compatible with viral replication (Fig. 2A). HDV RNA production was then induced with tetracycline to amplify functional or even ameliorated HDV genomes. Two days after induction, total RNA was extracted, reverse transcribed (RT) using random primers, and HDV cDNA was amplified by PCR. Because a 199 nt fragment from the right terminal region of genomic HDV, including ~60 nt of HDAg ORF, was previously reported to act as an RNA promoter for RNAP II (Beard et al., 1996; Abrahem and Pelchat, 2008), primers designed to specifically amplify this region were used (Fig. 1; gray rectangle), as described previously (Greco-Stewart et al., 2007; Abrahem and Pelchat, 2008). The sample quality was verified by agarose gel electophoresis, and the identity of the sequence was confirmed by Sanger sequencing (data not shown). To control for mutations introduced during either the RT-PCR or the deep-sequencing protocol, a genomic HDV RNA with the same sequence (hereinafter referred to as reference sequence) was synthesized by in vitro transcription with T7 RNAP and similarly processed. Both populations were tagged with a different bar code during the PCR for multiplexing, mixed at a ratio of about 1:100 control:viral population, and sent for deep-sequencing using the 454 Roche technology. We obtained 2510 and 747,158 readings for the control and the viral population, respectively.

Refinement of the populations

As reported previously, readings obtained by the 454 Roche technology usually include unrelated sequences, are heterogeneous in length and contain base calling errors (Gorzer et al., 2010; Beerenwinkel et al., 2012). Consequently, we developed a pipeline to refine the readings by performing several filtering steps. To remove readings unrelated to HDV, we calculated identity scores by comparing each reading to both polarities of the reference sequence using ClustalW (Thompson et al., 1994). Most of the readings were of the expected length (i.e. ~200 nt), and were ~75% identical to the reference sequence (Fig. 2B and C). For the viral population, there was also a smaller cluster of readings of approximately 64 nt, but most of these readings had low identities to HDV with large variations in their identity scores (Fig. 2C). Inspection of several sequences in this population revealed that they were chimeras composed of HDV and unidentified sequences, which is consistent with the generation of chimeras caused by

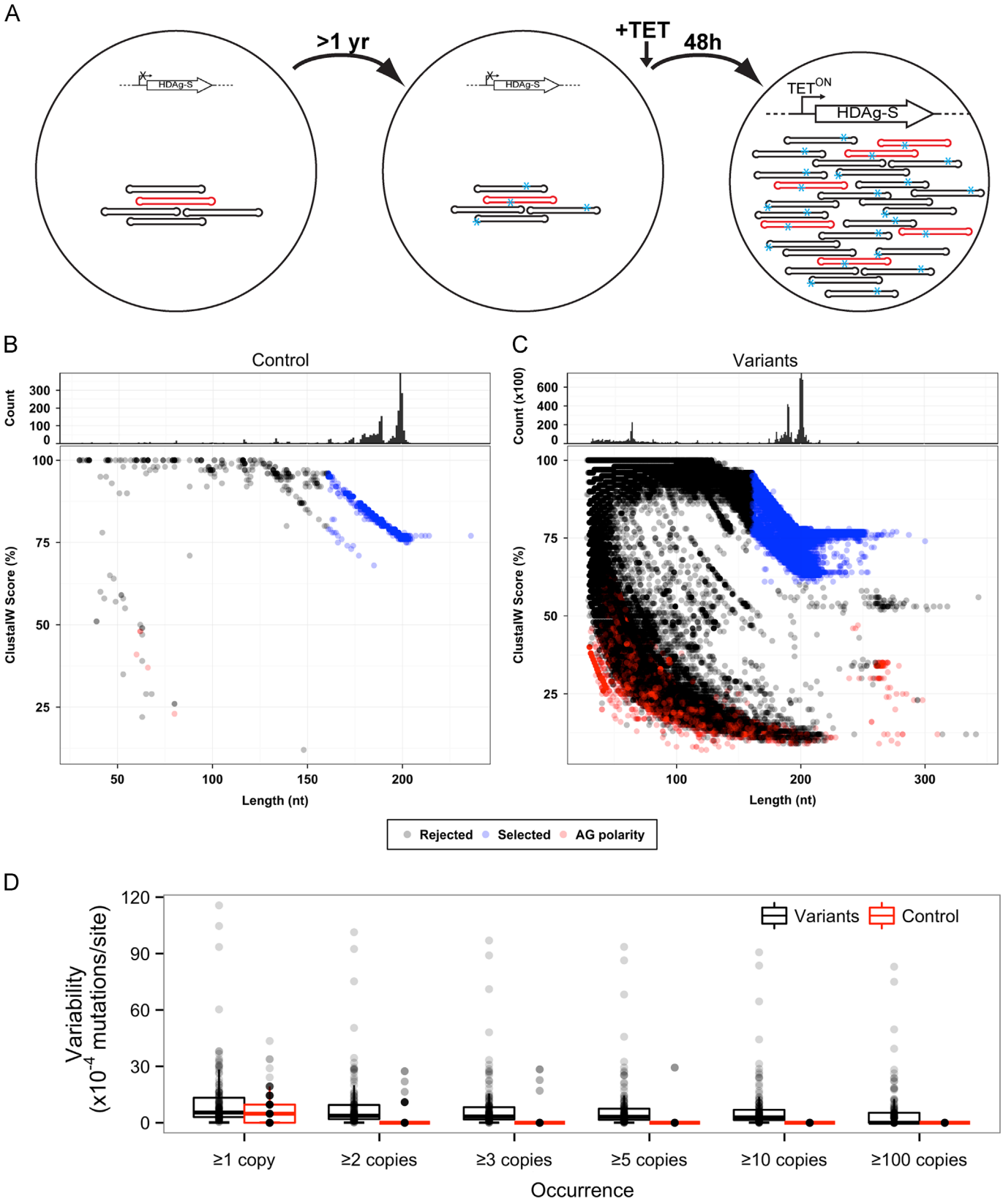

Fig. 2. Refinement of the sequence libraries generated by high-throughput sequencing of an HDV population. (A) Overview of the cellular system allowing HDV replication. A 293 cell line has been stably transfected with an HDV RNA genome containing a frame-shift deletion in the HDAg ORF and a plasmid allowing the production of the small antigen under the control of a promoter inducible by the addition of tetracycline. This cell line has been maintained for more than a year, allowing HDV replication at a basal level and the accumulation of mutations. Addition of tetracycline allowed HDV RNA production and amplification of functional or even ameliorated HDV genomes. (B) and (C) Filtering of the reading obtained by deep-sequencing using 454 technology according to the sequence length and percentage of identity to the reference sequence, for both the control (B) and the sequences amplified from 293-HDV cells (C). Top parts represent the number of sequences sorted according to their lengths. Bottom parts represent the percentage of identity of each reading to the reference sequence, as calculated by ClustalW (Thompson et al., 1994). Black and red dots indicate sequences with higher identities to the genomic and antigenomic polarity of the reference sequence, respectively. Sequences selected for further analysis are represented by blue dots. These sequences are at least 160 nt long and are at least 60% identical to the reference sequence. (D) Reduction of the background nucleotide variability by removal of sequences with low occurrence. Box plot representation of the nucleotide variability at each position calculated for sequences occurring at least 1, 2, 3, 5, 10 and 100 times. Black and red boxes indicate variants and control population, respectively.

non-specific amplification during deep-sequencing, as previously reported (Gorzer et al., 2010).

To obtain the information on the conservation of the secondary structure of this region, we only considered sequences of length longer than 160 nucleotides (Fig. 1; between nucleotides 1566 and 18) and at least 60% identical to the reference sequence (Fig. 2B and C; blue circles). Using these two constraints, 490,183 and 2070 sequences were selected for the viral population and the control, respectively. Furthermore, all of them were of genomic polarity, demonstrating the specificity of the primers used during amplification for this polarity of HDV RNA (Fig. 2B and C). Then, we performed pairwise alignment of each reading to the reference sequence, and calculated occurrences by clustering identical sequences (Fig. S1).

The cDNA amplification process and base calling error during deep-sequencing are known to generate apparent mutations that do not reflect selected variation (Beerenwinkel et al., 2012). To account for these errors, we calculated the nucleotide variability at each position, for both populations of sequences, and at different number of occurrences (Fig. 2D). Removal of sequences occurring less than three times greatly diminished nucleotide variability in the control without drastically affecting the number of different sequences in the viral population sample (Fig. 2D). Only seven positions varied in this subset of the control population, and none of these positions showed significant variability in the viral population (data not shown). The mutations in the control sample were likely generated by the additional PCR and/or transcription reaction used to generate the RNA species. Accordingly, only sequences occurring at least three times were kept for subsequent analysis. With these datasets, the overall nucleotide variations were 1.2×10^{-4} and 8.1×10^{-4} mutations/site for the control and the viral population, respectively. Altogether, we retained 1761 (85.07%) and 473,139 (96.52%) sequences representing 49 and 2351 different and recurring variants from the control and the viral population, respectively.

The right terminal region of genomic HDV RNA exists as a heterogeneous population in 293 cells

Based on the occurrence of the sequences, we found that one sequence was highly enriched and represented 76.3% of the variants obtained (i.e. 360,863 readings; Fig. S1). Interestingly, this sequence corresponded to the original HDV variant transfected into the cells (Chang et al., 2005). Despite of this, the number of different and recurring sequences obtained in the viral population sample suggested a larger sequence space generated during HDV replication. To evaluate the sequence space in our samples, neighbor-joining phylogenetic trees showing the genetic diversity of the different sequences composing both populations were generated. The trees were rooted on the reference sequence and plotted as circular dendrograms (Fig. 3). On the dendrograms, the size of the clusters in the trees is proportional to the occurrence of the sequences composing this cluster ($\log_2$ relationship). For the control, the clusters were phylogenetically close (Fig. 3, inset). The tree for the viral population from the 293 cells replicating HDV RNA was more heterogeneous, and numerous clusters with a high amount of sequences were phylogenetically distant from the reference sequence, indicating a larger sequence space. These results are in accordance with the accumulation of mutated variants during HDV replication, giving rise to a heterogeneous population (Wang et al., 1986; Chao et al., 1990, 1991).

Analysis of the variations found in the viral population

We calculated the nucleotide composition per position in order to determine the localization of position-specific variability and

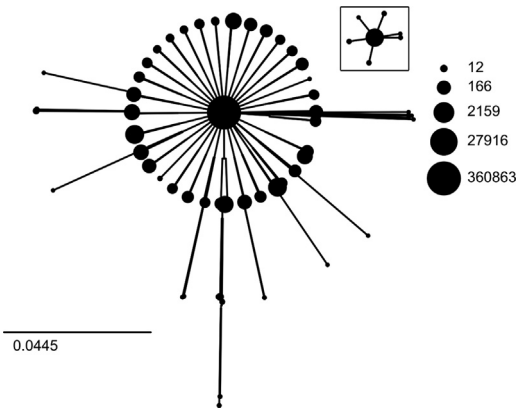

Fig. 3. Evaluation of the sequence space occupied by both the control population and the viral population replicating in 293 cells. Neighbor-joining phylogenetic trees rooted on the reference sequence were generated for both populations, and plotted as circular dendrograms. The sizes of the clusters are proportional to the occurrence of the sequences composing this cluster (examples of the $\log_2$ relationship between the surface of the circle and the number of sequences are found on the right side). The scale bar on bottom left indicates the distance as substitutions per site. The inset corresponds to the sequence space occupied by the control population, using the same scale as the viral population.

selective mutations. Variation was not homogeneous but varied according to the nucleotide position (Figs. 4 and S2). To discriminate between significant variability and background variations, we used four outlier tests (GESD, boxplot, medmad and shorth) to identify positions that appear to deviate from background variations (pooled as a "gray" zone on Fig. 4). Eleven positions showed significant variability in the four tests used, whereas the variability of 13 other positions was significant in at least one of the tests (Fig. 4; in red and blue, respectively). None of these 24 positions had significant variation in the control sequences. Because the variations were not homogeneous but fluctuated according to the nucleotide position, we recalculated the nucleotide variation rate. The variation rate of the viral population for these 24 positions was calculated to be 29.5×10^{-4} mutations/site, which is 24-fold the background variation rate calculated for the control (i.e. 1.2×10^{-4} mutations/site). For the other positions, the variation rate was 3.3×10^{-4} mutations/site, which is in the same order as background variation derived from the control. Noteworthily, the sequence located at the tip of the rod-like stucture (i.e. from position 1632 to 1557) was the most conserved. This conservation suggests that the sequence of this region might be required for the initiation to take place or for promoter recognition by the host transcription machinery. However, we cannot exclude the possibility that this motif might be associated with another activity unrelated to transcription initiation.

The analysis of the type of selected nucleotide changes revealed that 91.9% were transitions (either purine -> purine or pyrimidine -> pyrimidine) and 8.1% were transversions (purine -> pyrimidine or pyrimidine -> purine) (Fig. S2). Noteworthily, the highest nucleotide variation corresponded to a A->G transition located at position 1597. This selective mutation was observed in ~1% of the viral population (i.e. 4583 readings). This mutation is unlikely to be caused by experimental error since it was not found in the control population. This position is the second nucleotide of the anticodon CAU, which corresponds to the AUG initiation codon of HDAg on the HDV mRNA. Since HDAg-S, which is required for HDV replication, is provided *in trans* in the cellular system we used, a decrease of the selective pressure for the sequences able to produce the HDAg mRNA was expected. Interestingly, this mutation allows the conservation of the RNA secondary structure at this location by allowing Wobble base-pairing of the G with the U of the lower

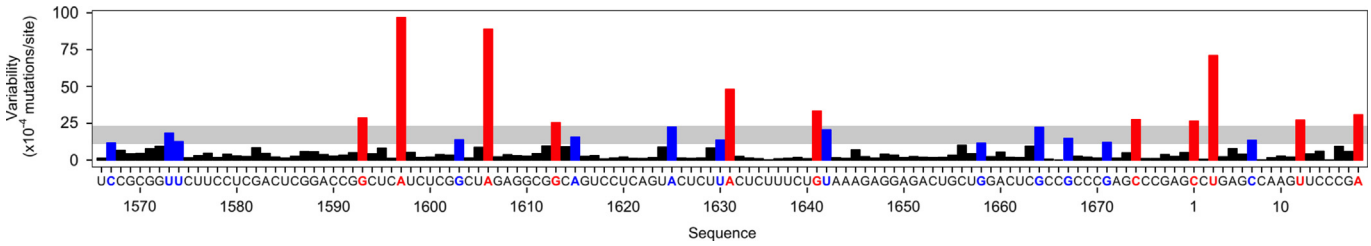

Fig. 4. Position-specific variability of the right terminal domain of genomic HDV RNA obtained from high-throughput sequencing. Representation of nucleotide variability for each position obtained from the refined alignment. The consensus sequence displayed corresponds to a region from nucleotides 1566 to 18 of the genomic polarity of HDV RNA. Four outlier tests were used to identify positions that appear to deviate from background variations, and their cut-offs was used to define the "gray zone". The calculated cut-offs were 15.58×10^{-4}, 23.33×10^{-4}, 10.79×10^{-4}, 11.19×10^{-4} mutations/site for GESD, boxplot, medmad and shorth, respectively. Blue and red indicate the position with significant variability in at least one or all four of the tests used, respectively. The U residue at location 1638 (gray on Fig. 5) was not used in the analysis due to high variability caused by the homopolymer effects during high-throughput 454 sequencing (Huse et al., 2007).

strand of genomic HDV RNA, suggesting the importance of the base pair at this location.

We next investigated conservation of the base pairs of this region. As a first step, we used the most energetically stable predicted secondary structure to assess base pair covariation. Then the secondary structure was manually adjusted based on the conservation derived from our dataset and is presented in Fig. 5A. The secondary structure derived from the conserved base pairs is also in accordance with a previously reported structure derived from in vitro nuclease mapping (Beard et al., 1996). We calculated the frequencies and compositions for each base pair across the alignment (Fig. S3). Sequences derived from the control allowed to establish a baseline for base pair variability of 2.1×10^{-4} mutations/pb (i.e. occurring 100 times). With this cut-off, we found that most of the nucleotide changes corresponded to transitions that enable the maintenance of the base pairs of either genomic or antigenomic polarities of HDV RNA (Fig. 5A). Interestingly, the majority of the variations are transitions generating G-U Wobble base pairs on antigenomic strand (Fig. 5A, CA and AC with a yellow background on genomic polarity), suggesting the importance of the secondary structure for this polarity. However, we cannot exclude the possibility that non-canonical C-A base pairs might also form on genomic strand.

Using the same approach, we calculated the frequencies and compositions of each bulge across the alignment occurring with variability of at least 2.1×10^{-4} mutations/site (i.e. occurring 100 times; Fig. 5B). We identified three bulges containing enriched mutations in their composition: U_{1599}/C_1C_2 -> U_{1599}/U_1C_2 (24.88×10^{-4} mutations/site; 1177 readings), A_{1606}/G_{1671} -> G_{1606}/G_{1671} (85.94×10^{-4} mutations/site; 4050 readings), and $U_{1629}U_{1630}A_{1631}$ -> $U_{1629}U_{1630}G_{1631}$ (46.01×10^{-4} mutations/site; 2177 readings). Our analysis also suggests the formation of a homopurine pair between A_{1606} and G_{1671}. Furthermore, the bulge at the initiation site always contains at least one uridine. Conservation of this uridine residue at this location is probably necessary for efficient initiation of complementary strand synthesis, since RNAP II is known to preferentially initiate transcription with purine residues (Baumann et al., 2010). Unfortunately, preliminary prediction of non-canonical base pairs within these bulges using the isostericity matrices was inconclusive due to the high conservation of this region (Leontis et al., 2002).

Conserved features of the right terminal stem-loop region of genomic HDV RNA in isolated variants

The previous analysis was performed on a viral population replicating in a specific cellular system. To assess the biological significance of our results and variations caused by the use of different hosts that might have different selection pressures, we analyzed both the positions of nucleotide conservation and covariation by extracting this region in sequences corresponding to HDV variants isolated from various hosts. Based on sequence analysis, the *Deltavirus* genus was previously classified into several major clades (Deny, 2006). We selected only the sequences from clade 1 since the viral population analyzed above was generated from a variant from this clade. Also, we decided to keep identical sequences in order to take into account selected sequence fitness. The sequences were extracted from the Subviral RNA database (Rocheleau and Pelchat, 2006), aligned using ClustalW (Thompson et al., 1994), and analyzed as above (Fig. S4).

Nucleotide comparison of the proposed initiation site for the transcription from this domain (i.e. nucleotide 1630; (Gudima et al., 1999; Abrahem and Pelchat, 2008)) and the surrounding nucleotides (i.e. nucleotides 1592 to 8 of the genomic polarity) in all the variants analyzed revealed a sequence heterogeneity pattern similar to what we observed in the viral population isolated from 293 cells. Specifically, the sequence is the most conserved at the tip of the rod-like structure with variation mostly upstream from the tip (Fig. 6A). We also calculated frequencies and compositions for each base pair across the alignment, as performed above (Fig. S5). The number of base pair variations was reduced as compared to the viral population in 293 cells, likely due to the small number of sequences available (i.e. 40 variants). Despite this, the majority of the variations are transitions allowing base pairs on either or both polarities of the HDV RNA genome, including generation of G-U Wobble base pairs on antigenomic strand (Fig. 6B, CA and AC with a yellow background on genomic polarity), analogous to our observation from the sequences derived from the viral population in 293 cells. In addition, several of the mutations observed in the bulges derived from the viral population were also found in these clade 1 variants, including the enrichment of purines at position 1606 and 1671 (Fig. 6B, blue rectangles).

Discussion

Previous studies indicated a role for the right terminal region of genomic HDV RNA in viral replication. This region includes the site of transcription initiation for HDAg mRNA, binds an active RNAP II pre-initiation complex, acts as template to initiate antigenomic RNA synthesis in vitro, and mutations affecting the rod-like conformation of this region decrease both RNAP II affinity/initiation and HDV RNA accumulation in cells (Beard et al., 1996; Gudima et al., 1999; Greco-Stewart et al., 2007; Abrahem and Pelchat, 2008). Here, we took advantage of next-generation sequencing technology to generate 473,139 new HDV sequences of this region from a cellular system in which the selective pressure was mainly on viral replication.

Because HDAg-S could not be produced by the mutated HDV RNA genome in this replication system, but provided *in trans* by the cells, we were expecting reduced selective pressure on both

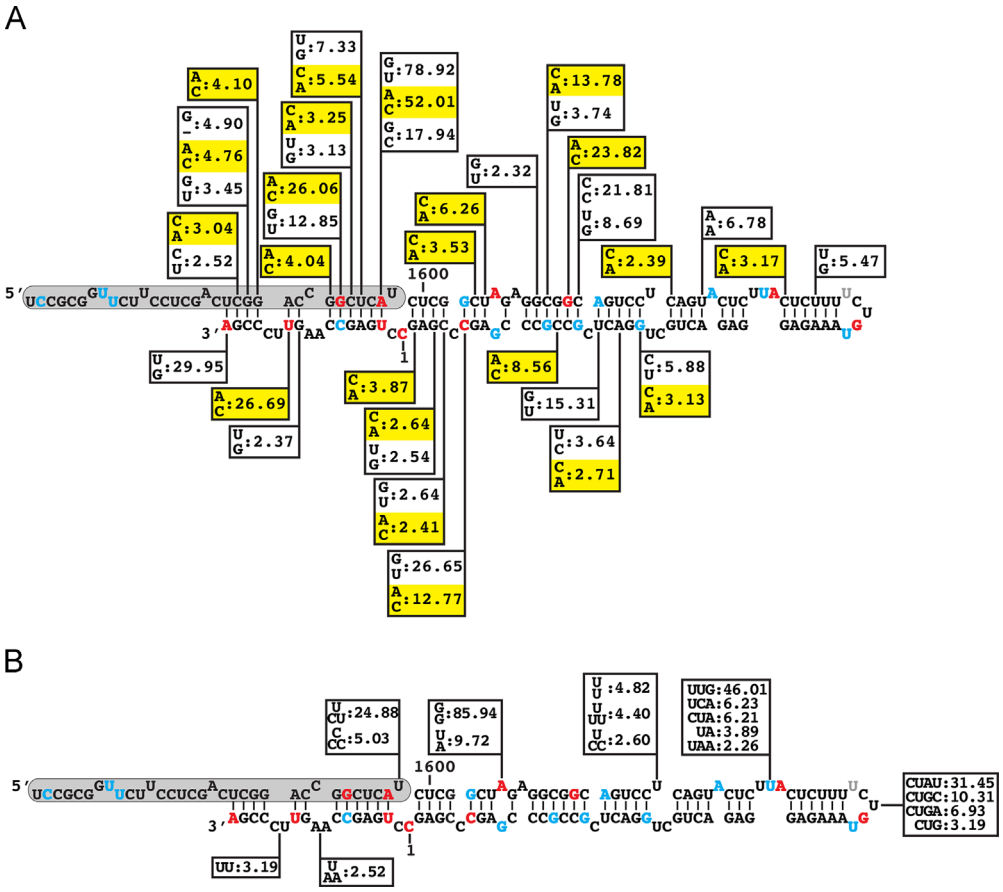

Fig. 5. Covariation analysis of the right terminal domain of genomic HDV RNA obtained from high-throughput sequencing. The covariation variability of every base pairs (A) and single-stranded region (B) was calculated and displayed on the consensus RNA secondary structure. Blue and red nucleotides indicate the position with significant variability, as determined in Fig. 4. Yellow background indicates transitions generating C-A on genomic HDV RNA. The gray rectangles represent the 5′-end of HDAg ORF. All numbers correspond to $\times 10^{-4}$ mutations/site. The gray U residue at location 1638 was not used in the analysis due to high variability caused by the homopolymer effects during high-throughput 454 sequencing (Huse et al., 2007).

the sequence and the secondary structure of the region corresponding to either the HDAg ORF or its promoter. However, our analysis indicates that both the sequence and the secondary structure of this segment of HDV are very conserved. This suggests that, in addition to its proposed role as promoter for HDAg transcription, this region is also involved in HDV replication. The highest nucleotide variation corresponded to a A- > G transition at position 1597, which was observed in ~1% of the viral population. This mutation disrupts the initiation codon of HDAg but is still predicted to allow base pairing with the opposite strand. The importance of a base pair at this location is in agreement with a recent study reporting that mutations disrupting the base pairing at this location hinder HDV replication (Liao et al., 2012).

Analysis of the sequences corresponding to the right terminal region of genomic HDV RNA revealed that this region is less heterogeneous than expected based on previous reports on isolated HDV variants (Wang et al., 1986; Chao et al., 1990, 1991). One of the sequences was highly enriched and represented 76.3% of the variants obtained (i.e. 360,863 readings), suggesting enhanced fitness for this sequence, which also corresponded to the original HDV variant transfected into the cells (Chang et al., 2005). Despite this, comparison of the sequence space between the viral and the control populations revealed that this region of the HDV RNA genome is heterogeneous in 293 cells, consistent with the notion that HDV RNA forms a population of different sequences due to the infidelity of a "DNA-dependent" RNAP acting on an RNA template (Wang et al., 1986; Chao et al., 1990, 1991). In total, 2351 different and recurring sequences were found in the sample derived from the viral population. However, due to the approach used, we were not able to distinguish between variations that occurred over the year of replication from those following HDAg-S induction by tetracycline. Interestingly, nucleotide variations for the viral population were not distributed evenly and 24 positions with higher variability were identified. The variation rate of these "hot spots" was calculated to be 29.5×10^{-4} mutations/site and account for the larger sequence space observed for the viral population. We cannot completely exclude the possibility that some of the sequence diversity observed might have been artificially generated during the protocols used. However, it is unlikely due to the short length of the cDNA fragment, and because throughout our pipeline, we used a control sample of the same sequence to establish cut-offs to account for the error-rate due to the experimental steps of the reverse-transcription, the PCR amplification and the deep-sequencing. Based on the control, we calculated that an overall variation rate of 1.2×10^{-4} mutations/site might be due to the protocols used.

Although it reflects the mutation rate during HDV replication in a non-physiological cellular system (i.e. 293 cells) where the antigen is provided in *trans*, our calculated variation rate for the viral population (i.e. overall 8.1×10^{-4} mutations/site) is in accordance with mutation rates calculated for other RNA viruses (i.e. 10^{-3}–10^{-5} substitutions/nt), which have a high polymerase error-rate, giving rise to a heterogeneous population (Domingo et al., 2012). It is also one order of magnitude higher than what is reported for RNAP II when acting on DNA templates (i.e. 10^{-5} substitutions/nt; (Cramer, 2004)). This suggests that RNAP II has

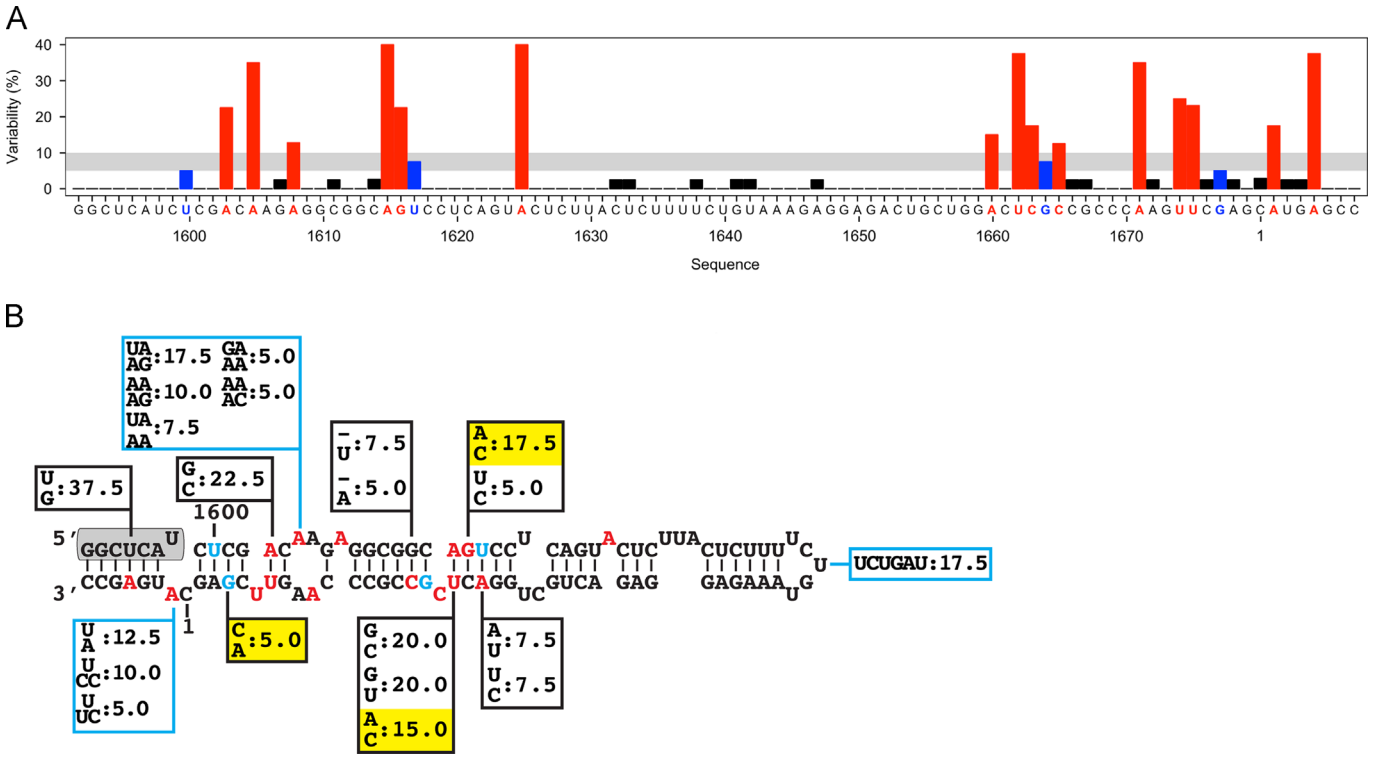

Fig. 6. Conserved RNA features of the right terminal region of genomic HDV from clade I variants. (A) Representation of nucleotide variability for each position obtained from the secondary structure alignment of 40 clade I HDV variants extracted from the Subviral RNA Database (Rocheleau and Pelchat, 2006). The consensus sequence displayed corresponds to a region from nucleotides 1592 to 8 of the genomic polarity of HDV RNA. (B) The covariation variability of every base pair was calculated and displayed on the consensus RNA secondary structure. Yellow background indicates transitions generating C-A on HDV RNA of genomic polarity. The gray rectangles represent the 5′-end of HDAg ORF. The blue rectangles indicate covariation variabilities in single-stranded regions. All numbers correspond to percentage of mutations/site. Blue and red indicate the position with variability of at least 5% and 10%, respectively.

an increased mutation rate when acting on HDV RNA, which is also supported by studies showing that HDAg-S might accelerate forward translocation of the polymerase at the cost of fidelity (Yamaguchi et al., 2001). However, we cannot exclude the possibility that the observed variations might also be due to the activity of another protein. Previous analysis of the sequences of a few HDV RNA genome variants in the same cellular system estimated a variation rate of 2.1% changes/nucleotide/year reported for the complete HDV RNA genome, and attributed most of the mutations to adenosine deaminase acting on RNA (ADAR) activity (Chang et al., 2005). Interestingly, in this study, the sequence of the right terminal region was conserved. Our estimated variation rate is also lower than what was reported on complete HDV RNA sequences from sequential isolates (i.e. $2–3 \times 10^{-3}$ changes/nucleotide/year; (Weiner et al., 1988; Lee et al., 1992)), and what was calculated for viroids, small single-stranded circular RNA genomes similar to HDV but replicating in plants (i.e. 2.5×10^{-3} changes/site/replication cycle; (Gago et al., 2009)). However, the mutation frequencies calculated for viroids were derived from samples isolated in the context of a natural infection, with more selective pressure from their host.

More importantly, we found that the sequence at the tip of the rod-like structure of this region is very conserved in both sequences derived from the viral population in 293 cells (Fig. 4) and clade 1 variants (Fig. 6A). This region is composed of a stretch of pyrimidines upstream of the terminal loop, which is matched on the opposite strand by a region containing almost exclusively purines, allowing the conservation of the rod-like structure of this region. This is consistent with previous reports on both a decrease of HDV accumulation in cells and reduced RNAP II interaction by the inversion of the strands of the tip region (i.e. “flip” mutant), suggesting that in addition to the structure, sequence conservation is important for viral replication (Wu et al., 1997; Greco-Stewart et al., 2007; Abrahem and Pelchat, 2008). It is also possible that the sequence of either or both strands serves as a binding site for either polypurine or polypyrimidine binding proteins. One candidate protein is PSF, a polypurine binding protein we recently reported to bind this region, and which is also known to associate with RNAP II (Emili et al., 2002; Greco-Stewart et al., 2006).

Most of the nucleotide changes were upstream of the tip and our results indicate that they were selected to maintain the rod-shaped secondary structure of either polarity of this region of HDV RNA. This suggests that the secondary structure of these regions is important for HDV replication/transcription and is in accordance with previous experiments in which mutagenesis disturbing the secondary structure of this region affected both HDV accumulation in cells and RNAP II binding (Beard et al., 1996; Wu et al., 1997; Gudima et al., 1999; Greco-Stewart et al., 2007; Abrahem and Pelchat, 2008). Although, several specific nucleotides within this region were reported to be essential for high level of HDV accumulation in cells, in most of the cases the mutations introduced could also disrupt base pairing. Interestingly, our high-throughput sequencing of this region indicates that most of the selected nucleotide changes corresponded to transitions to maintain the secondary structure of the antigenomic polarity (i.e. generating G-U on antigenomic strand), suggesting the involvement of the antigenomic strand of this region in HDV replication. In support of this hypothesis, both strands of this region associate with RNAP II (Greco-Stewart et al., 2007), and a small 5′-capped HDV RNA of genomic polarity corresponding to this region was identified during a screening for small RNAs in cells replicating HDV (Haussecker et al., 2008). This small HDV RNA might represent a transcription product from antigenomic RNA. Additionally, an HDV cDNA fragment corresponding to this region was

reported to have bidirectional promoter activity, although it was never confirmed using RNA fragments (Macnaughton et al., 1993; Liao et al., 2012).

Conclusion

In this study, we used next-generation sequencing technology to define selected features on the right terminal domain of genomic HDV RNA from HDV variants isolated in cellular system in which the selective pressure was mainly on viral replication. We analyzed both the sequences and the secondary structural implication by investigating nucleotide conservation and positions of covariation in a dataset composed of 473,139 sequences representing 2351 new HDV variants for this region. We also corroborated our finding with sequences from HDV variants isolated in various hosts. Our analysis suggests a precise RNA secondary structure for this region and indicates the conservation of the nucleotides at the tip of the rod-like structure. Both features might be important for HDV replication, likely through the recruitment of RNAP II, and are in accordance with previous mutagenesis on this region of the HDV RNA genome. We also developed a pipeline to filter, align and analyze the sequences, which could be a useful strategy for future high-throughput sequencing analysis of sequence conservation and base pair co-variation in an RNA population. More importantly, the selected features identified in this study might be useful in identifying other RNA promoters for RNAP II, including in human RNAs.

Materials and methods

Cell culture and HDAg-S induction

The 293-HDV cells are 293 cells stably transfected with a plasmid encoding HDAg-S under the control of tetracycline and an HDV RNA genome deficient in HDAg-S production, and were kindly provided by John Taylor (Chang et al., 2005). The 293-HDV cells were grown at 37 °C with 5% CO_2 in DMEM supplemented with 10% calf serum (CS), hygromycin and blasticidin. Viral replication was induced upon addition of 1 μg/ml of tetracycline, and two days later the total RNA was extracted with Trizol (Invitrogen) according to the manufacturer's recommendations.

In vitro transcription of the control population

To serve as a control population to account for mutations introduced during the protocols used, a genomic HDV RNA, with the same sequence as the variant originally transfected into the cellular system used (Chang et al., 2005), was synthesized by in vitro run-off transcription using T7 RNAP (New England Biolabs; Pickering, Ontario, Canada; NEB), as previously described (Greco-Stewart et al., 2007; Abrahem and Pelchat, 2008). To generate the cDNA for the transcription reaction, PCR amplification was performed on a plasmid encoding a dimer of the HDV genome (pHDVd2) with both sense (5′–GAATTCTAATACGACTCACTATAGGG^{1541}ACTGCTCGAGGATCTCTTCTCTCC^{1564}-3′; underlined nucleotide sequence indicates T7 promoter) and antisense (5′-^{60}ACATCCCCTCTCGGGTAC^{43}-3′) oligonucleotides. After transcription, the DNA template was digested with DNase I (NEB) for 30 min at 37 °C and the RNA was fractionated by 7M urea denaturing polyacrylamide gel electrophoresis (PAGE) in 1XTBE buffer (100 mM Tris–borate, pH 8.3, 1 mM EDTA). The band corresponding to the control RNA was visualized by UV shadowing, excised, and eluted overnight in 500 mM ammonium acetate, 0.1% SDS. The RNA was then precipitated in ethanol, resuspended in H_2O, desalted by Sephadex G-50 columns (GE Healthcare), and precipitated in ethanol. The purified control RNA was resuspended in H_2O, quantified by spectrophotometry at 260 nm and stored at −20 °C.

Reverse-transcription and PCR

Both the control population and total RNA from 293-HDV cells were reverse transcribed with random primers according to the manufacturer's instructions (Biorad). cDNAs were then used as templates for PCR amplifications with Deep Vent polymerase (NEB). For both cDNAs, the antisense primer used was 5′–CCTATCCCCTGTGTGCCTTGGCAGTCTCAG^{54}CCTCTCGGGTACTGATCCTCCCCCCGCGTCTCCTCG^{19}-3′. The sense primers were C∗C∗A∗T∗CTCATCCCTGCGTGTCTCCGACTCAGACGAGT∗G∗C∗G∗† T^{1541}ACTGCTCGAGGATCTCTTCTCTCCC^{1565}-3′ and C∗C∗A∗T∗CTCATCCCTGCGTGTCTCCGACTCAGATCAGA∗C∗A∗C∗G^{1541}ACTGCTCGAGGATCTCTTCTCTCCC^{1565}-3′ for the control and the viral population, respectively (the ∗ indicate phosphorothioate modifications). The PCR products were purified from a 1.5% agarose gel (Qiagen), and the identity of the cDNAs was confirmed by Sanger sequencing (StemCore facilities, Ottawa Hospital Research Institute). One microgram of the viral population DNA and 10 ng of the control DNA were pooled, and sent for deep-sequencing using the Roche 454 GS FLX Titanium platform (McGill sequencing facilities, Genome Quebec). Raw sequencing data from both samples were deposited on the Sequence Read Archive of NCBI [SRA: SRR765851, SRR765852].

Analysis of HDV variants from deep-sequencing

For each sequence, the name of the sequence, the composition in nucleotides, the length and the sequencing quality score were stored in a database. The percentage of identity of each sequence to both polarities of the reference HDV sequence was calculated with ClustalW 2.1 (Thompson et al., 1994) and stored in the database. A cut-off of 160 nt of length and 60% identity was used to select the sequences for alignment with Mosaik 1-3.0 from the Marth laboratory (http://bioinformatics.bc.edu/marthlab/Mosaik). In house Perl-scripts were used to cluster the sequences based on identity, and to obtain statistics on nucleotide composition. In house R-scripts were used to analyze the correlation between the variability of the sequences and the number of identical sequences, and analyze both nucleotide composition and covariation. An in house R-script was used to detect hot spots of variability by using cutoff generated by selecting both the minimum and maximum of four outlier detection procedures included in the R package Parody (i.e. ("GESD", "boxplot", "medmad" and "shorth") (http://www.bioconductor.org/packages/release/bioc/html/parody.html). Neighbor-joining phylogeny of the sequences was performed with the R package APE (Thompson et al., 1994; Paradis et al., 2004), and the trees were drawn using a modified radial.phylog R-script from the package ADE4 (Dray and Dufour, 2007). Secondary structure prediction was performed with Mfold (Zuker, 2003).

Analysis of HDV variants from various hosts

The HDV sequences were taken from the Subviral RNA Database (http://subviral.med.uottawa.ca/; (Rocheleau and Pelchat, 2006)). The sequences were first aligned with ClustalW 2.1 (Thompson et al., 1994) and neighbor-joining phylogeny of the sequences was performed with the R package APE (Thompson et al., 1994; Paradis et al., 2004). The sequences clustering with known clade 1 variants were extracted and realigned with ClustalW 2.1. In house R- and PERL-scripts were used to analyze both the composition and nucleotide variation from the alignment, as performed with the HDV sequences generated from deep-sequencing.

Acknowledgments

This work was funded by a grant from the Natural Science and Engineering Research Council of Canada (NSERC Canada) awarded to M. Pelchat. The authors wish to acknowledge Alfredo Staffa of the genotyping platform of the McGill University and Genome Quebec Innovation Centre for his technical assistance. The authors would like to thank Professor Earl G. Brown, Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, Canada, for giving valuable suggestions.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2013.12.017.

References

Abrahem, A., Pelchat, M., 2008. Formation of an RNA polymerase II preinitiation complex on an RNA promoter derived from the hepatitis delta virus RNA genome. Nucleic Acids Res. 36, 5201–5211.
Baumann, M., Pontiller, J., Ernst, W., 2010. Structure and basal transcription complex of RNA polymerase II core promoters in the mammalian genome: an overview. Mol. Biotechnol. 45, 241–247.
Beard, M.R., MacNaughton, T.B., Gowans, E.J., 1996. Identification and characterization of a hepatitis delta virus RNA transcriptional promoter. J. Virol. 70, 4986–4995.
Beerenwinkel, N., Gunthard, H.F., Roth, V., Metzner, K.J., 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front Microbiol. 3, 329.
Borderia, A.V., Stapleford, K.A., Vignuzzi, M., 2011. RNA virus population diversity: implications for inter-species transmission. Curr. Opin. Virol. 1, 643–648.
Casey, J.L., 2012. Control of ADAR1 editing of hepatitis delta virus RNAs. Curr. Top. Microbiol. Immunol. 353, 123–143.
Chang, F.L., Chen, P.J., Tu, S.J., Wang, C.J., Chen, D.S., 1991. The large form of hepatitis delta antigen is crucial for assembly of hepatitis delta virus. Proc. Natl. Acad. Sci. U.S.A 88, 8490–8494.
Chang, J., Gudima, S.O., Tarn, C., Nie, X., Taylor, J.M., 2005. Development of a novel system to study hepatitis delta virus genome replication. J. Virol. 79, 8182–8188.
Chang, J., Nie, X., Chang, H.E., Han, Z., Taylor, J., 2008. Transcription of hepatitis delta virus RNA by RNA polymerase II. J. Virol. 82, 1118–1127.
Chao, Y.C., Chang, M.F., Gust, I., Lai, M.M., 1990. Sequence conservation and divergence of hepatitis delta virus RNA. Virology 178, 384–392.
Chao, Y.C., Lee, C.M., Tang, H.S., Govindarajan, S., Lai, M.M., 1991. Molecular cloning and characterization of an isolate of hepatitis delta virus from Taiwan. Hepatology 13, 345–352.
Cramer, P., 2004. Structure and function of RNA polymerase II. Adv. Protein Chem. 67, 1–42.
Deny, P., 2006. Hepatitis delta virus genetic variability: from genotypes I, II, III to eight major clades? Curr. Top. Microbiol. Immunol. 307, 151–171.
Domingo, E., Sheldon, J., Perales, C., 2012. Viral quasispecies evolution. Microbiol. Mol. Biol. Rev. 76, 159–216.
Dray, S., Dufour, A.B., 2007. The ade4 package: implementing the duality diagram for ecologists. J. Stat. Softw. 22, 1–20.
Emili, A., Shales, M., McCracken, S., Xie, W., Tucker, P.W., Kobayashi, R., Blencowe, B. J., Ingles, C.J., 2002. Splicing and transcription-associated proteins PSF and p54nrb/nonO bind to the RNA polymerase II CTD. RNA 8, 1102–1111.
Filipovska, J., Konarska, M.M., 2000. Specific HDV RNA-templated transcription by pol II in vitro. RNA 6, 41–54.
Fu, T.B., Taylor, J., 1993. The RNAs of hepatitis *delta* virus are copied by RNA polymerase II in nuclear homogenates. J. Virol. 67, 6965–6972.
Gago, S., Elena, S.F., Flores, R., Sanjuan, R., 2009. Extremely high mutation rate of a hammerhead viroid. Science 323, 1308.
Gorzer, I., Guelly, C., Trajanoski, S., Puchhammer-Stockl, E., 2010. The impact of PCR-generated recombination on diversity estimation of mixed viral populations by deep sequencing. J. Virol Methods 169, 248–252.
Greco-Stewart, V.S., Miron, P., Abrahem, A., Pelchat, M., 2007. The human RNA polymerase II interacts with the terminal stem-loop regions of the hepatitis *delta* virus RNA genome. Virology 357, 68–78.
Greco-Stewart, V.S., Thibault, C.S., Pelchat, M., 2006. Binding of the polypyrimidine tract-binding protein-associated splicing factor (PSF) to the hepatitis delta virus RNA. Virology 356, 35–44.
Gudima, S., Dingle, K., Wu, T.T., Moraleda, G., Taylor, J., 1999. Characterization of the 5′ ends for polyadenylated RNAs synthesized during the replication of hepatitis *delta* virus. J. Virol. 73, 6533–6539.
Gudima, S., Wu, S.Y., Chiang, C.M., Moraleda, G., Taylor, J., 2000. Origin of hepatitis delta virus mRNA. J. Virol. 74, 7204–7210.
Haussecker, D., Cao, D., Huang, Y., Parameswaran, P., Fire, A.Z., Kay, M.A., 2008. Capped small RNAs and MOV10 in human hepatitis delta virus replication. Nat. Struct. Mol. Biol. 15, 714–721.
Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., Welch, D.M., 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 8, R143.
Kuo, M.Y., Chao, M., Taylor, J., 1989. Initiation of replication of the human hepatitis delta virus genome from cloned DNA: role of delta antigen. J. Virol. 63, 1945–1950.
Kuo, M.Y., Goldberg, J., Coates, L., Mason, W., Gerin, J., Taylor, J., 1988. Molecular cloning of hepatitis delta virus RNA from an infected woodchuck liver: sequence, structure, and applications. J. Virol. 62, 1855–1861.
Lee, C.M., Bih, F.Y., Chao, Y.C., Govindarajan, S., Lai, M.M., 1992. Evolution of hepatitis delta virus RNA during chronic infection. Virology 188, 265–273.
Leontis, N.B., Stombaugh, J., Westhof, E., 2002. The non-Watson-Crick base pairs and their associated isostericity matrices. Nucleic Acids Res. 30, 3497–3531.
Liao, F.T., Hsu, L.S., Ko, J.L., Lin, C.C., Sheu, G.T., 2012. Multiple genomic sequences of hepatitis delta virus are associated with cDNA promoter activity and RNA double rolling-circle replication. J. Gen. Virol. 93, 577–587.
Macnaughton, T.B., Beard, M.R., Chao, M., Gowans, E.J., Lai, M.M., 1993. Endogenous promoters can direct the transcription of hepatitis delta virus RNA from a recircularized cDNA template. Virology 196, 629–636.
MacNaughton, T.B., Gowans, E.J., McNamara, S.P., Burrell, C.J., 1991. Hepatitis *delta* antigen is necessary for access of hepatitis *delta* virus RNA to the cell transcriptional machinery but is not part of the transcriptional complex. Virology 184, 387–390.
Moraleda, G., Taylor, J., 2001. Host RNA polymerase requirements for transcription of the human hepatitis delta virus genome. J. Virol. 75, 10161–10169.
Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.
Rocheleau, L., Pelchat, M., 2006. The Subviral RNA Database: a toolbox for viroids, the hepatitis delta virus and satellite RNAs research. BMC Microbiol. 6, 24.
Ryu, W.S., Bayer, M., Taylor, J., 1992. Assembly of hepatitis delta virus particles. J. Virol. 66, 2310–2315.
Sureau, C., Moriarty, A.M., Thornton, G.B., Lanford, R.E., 1992. Production of infectious hepatitis delta virus in vitro and neutralization with antibodies directed against hepatitis B virus pre-S antigens. J. Virol. 66, 1241–1245.
Taylor, J.M., 2009. Replication of the hepatitis delta virus RNA genome (Chapter 3). Adv. Virus Res. 74, 103–121.
Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.
Wang, K.S., Choo, Q.L., Weiner, A.J., Ou, J.H., Najarian, R.C., Thayer, R.M., Mullenbach, G.T., Denniston, K.J., Gerin, J.L., Houghton, M., 1986. Structure, sequence and expression of the hepatitis delta (delta) viral genome. Nature 323, 508–514.
Weiner, A.J., Choo, Q.L., Wang, K.S., Govindarajan, S., Redeker, A.G., Gerin, J.L., Houghton, M., 1988. A single antigenomic open reading frame of the hepatitis delta virus encodes the epitope(s) of both hepatitis delta antigen polypeptides p24 delta and p27 delta. J. Virol. 62, 594–599.
Wu, T.T., Netter, H.J., Lazinski, D.W., Taylor, J.M., 1997. Effects of nucleotide changes on the ability of hepatitis delta virus to transcribe, process, and accumulate unit-length, circular RNA. J. Virol. 71, 5408–5414.
Yamaguchi, Y., Filipovska, J., Yano, K., Furuya, A., Inukai, N., Narita, T., Wada, T., Sugimoto, S., Konarska, M.M., Handa, H., 2001. Stimulation of RNA polymerase II elongation by hepatitis delta antigen. Science 293, 124–127.
Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31, 3406–3415.